

# Fast reconstruction of compact context-specific metabolic networks via integration of microarray data

Maria Pires Pacheco<sup>1</sup> and Thomas Sauter<sup>1</sup>

1 Life Sciences Research Unit, University of Luxembourg, Luxembourg \* E-mail: thomas.sauter@uni.lu

## 1 Introduction

Metabolism is a dynamic process that involves transport of metabolites and thousands of chemical reactions in which thousands of compounds are converted into others. Alternative pathways and branches are continuously activated or shut down to maximize metabolic efficiency in a specific context [18]. Metabolism is so complex that the underlying processes can hardly be understood without using simplified mathematical representations. The most comprehensive formulations are genome-scale Reconstructions (GEMs). For *homo sapiens* there are several of these reconstructions like Recon 1 and 2 [6,20] or the Edinburgh human metabolic network [11]. Alongside these reconstructions follows the development of extensive reaction databases, like the HMR [1,2] or HumanCyc [5,17], which collect information to refine the available models. EMs and models derived from GEMs following omics data integrations were successively used to understand how perturbations in the metabolism lead to severe pathologies targets [2,7,12].

GEMs are generic representations of a cell of an organism comprising all the reactions that can potentially get active regardless of the environment and cell type. Therefore they do not cover the fact that the set of expressed genes and thereby the set of active reactions vary significantly in function of the cellular context. This necessitates the generation of context-specific models containing only pathways predicted to be active in a given environment. Most context-specific reconstruction methods assume that the expression of genes correlates with the active state of the related reactions. Although, this assumption is only partially justified, context-specific models showed a higher predictive power than the GEMs from which they were derived from [3,9]. This is due to GEMs containing multiple alternative pathways that are rarely simultaneously active and therefore tend to have an increased number of false negatives in gene essentiality assays compared to context-specific models that only comprise the active alternative pathway [4,7,8].

Recently we proposed an algorithm for the fast reconstruction of compact context-specific metabolic networks (FASTCORE) that allowed dropping the reconstruction time of context-specific networks to the time order of seconds [21]. This extremely low computational demand opens new possibilities for improving the quality of the models. Several rounds of model reconstruction, testing of the models predictions against real experimental data, curation steps of the input model and the set of core reactions as well as cross-validations assays are required to reconstruct high-quality models. These semi-automated model curations steps are in such extend not possible with competing algorithms due to their high computational demands. FASTCORE requires as input a GEM and a set of core reactions being active in the context of interest. FASTCORE identifies a close to minimal set of non-core reactions from the input model to be added to the core set in order to obtain a consistent model, in which every reaction in the model is able to carry a non-zero flux. To reconstruct compact models, the inclusion of non-core reactions is penalized [21].

But the question which genes are expressed in a given cell-type and therefore the establishment of the core set of reactions is non-trivial. Microarray expression data, so far the most popular data source for model contextualization, allow comparing the probe expression levels between two conditions. But probe effects do not allow a direct comparison between the probe sets, as the amount of noise is non-negligible and varies within probe sets so that higher in-

tensity does not imply higher expression levels or even to state that a gene is expressed [23]. Further the intensities retrieved from microarray data are continuous while FASTCORE, like most context-specific algorithm require a binary input data (the establishing core set).

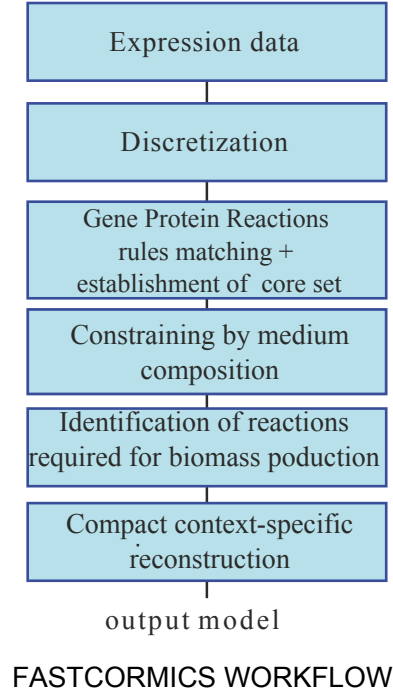
To adapt FASTCORE for the integration of microarray data, we therefore propose a new workflow: FASTCORMICS. FASTCORMICS requires as input microarray data and a GEM. Like FASTCORE, FASTCORMICS is devoid of heuristic parameter settings and has a low computational demand with overall building times in the order of a few minutes FASTCORMICS preprocesses the microarrays data with the discretization tool Barcode [14, 23]. Barcode uses prior knowledge on the intensity distribution of each probe set for a given microarray platform to segregate between expressed genes and non-expressed genes. The preprocessing step with Barcode allows circumventing the need of setting a heuristic expression threshold that segregates between expressed and non-expressed genes as e.g. in [3, 7, 24]. Choosing such a threshold is arbitrary and critical for the output metabolic models as in response to this threshold complete branches, alternative pathways, or subsystems might be included or excluded, thereby heavily changing the functionalities of the model. Further, Barcode shows a better correlation between predicted expression and protein expression than competing discretization methods for the segregation of gene expression and allows to reduce batch and lab-effects that affect measurements [23].

FASTCORMICS was validated via an essentiality assay performed on two cancer models (cancer1 and cancer2) extracted respectively from Recon 1 and Recon 2 by the FASTCORMICS workflow. The predicted essential genes were compared to a list of essential genes established in [10] in a shRNA knockdown screen on cancer cell lines. The predictive power of the reconstructed context-specific models was compared to the original GEMs and to a generic cancer model [7] built by the competing MBA algorithm [9]. Furthermore, as a second quality control step, enrichment in neoplasia-related genes retrieved from the DisGeNET database in the predicted essential genes was assessed with a hypergeometric test to check for the model ability to predict cancer relevant genes. In general, FASTCORMICS outperforms competing algorithms and allows obtaining high-quality, robust models in a high-throughput manner. This will allow the use of metabolic modelling as routine process for the analysis of microarray data e.g. in the field of personalized medicine.

## 2 Materials and methods

### 2.1 Workflow overview

The general workflow of FASTCORMICS (figure1) contains a discretization step with Barcode to obtain a list of genes expressed in the context of interest. The latter is then mapped to the consistent generic model via the model’s Gene-Protein-Reactions Rules (GPR) to obtain a list of active reactions (core reactions). For reactions that are under the control of one gene, the discretized gene expression value is directly mapped to the reaction; otherwise if more genes are associated to a reaction, the relationship between the genes and the reaction is given by Boolean Rules. A Boolean AND means that all the genes have to be expressed to activate the reaction, which is typically the case when a reaction is controlled by a complex of proteins. Therefore the minimum of the discretized values is mapped to the reaction. A Boolean OR signifies that only one gene has to be expressed, the maximal value is mapped to the reaction. Boolean ANDs and ORs can be combined inside a same rule i.e.  $((A \text{ AND } B) \text{ OR } C)$ , in this example the minimal value D is computed between A and B then the maximum between D and C is matched to the reaction. Reactions that are predicted active according to the GPR rules constitute the set of core reactions that are fed into a modified version of FASTCORE (mFC) that allows leaving



**Figure 1. FASTCORMICS Workflow:** After discretization of the microarray data with Barcode, the expressed genes are mapped to the input model according to the Gene-Protein-Reactions rules. The FASTCORE core set is composed of reactions under the control of Barcode-supported genes. Optionally, the model can be constrained in function of the medium composition and a biomass function. A modified version of FASTCORE, that allows the definition of a set of non-penalized reactions (in this study: barcode-supported core reactions) is run. The modified version of FASTCORE forces the biomass function to carry a non-zero flux while penalizing the inclusion of non-core reactions. The output of the modified FASTCORE is then added to the core set and the modified FASTCORE is run again, this time, forcing all core reactions to carry a flux while penalizing non-core reactions. Transporters are removed from the core set, but are not penalized as explained in the main text.

a set of reactions not penalized besides defining core and non-core reactions. The inclusion of the set of non-penalized reactions are, unlike core reactions, not forced but only preferred over the inclusion of non-core reactions, which are penalized. Barcode-supported transporters, are a good example for the need of this new reaction set. Transporter reactions are generally under the control of promiscuous genes (in the consistent version of Recon 2 [20], e.g. the gene SLC7A6 controls 294 reactions) and therefore transporters should be removed from the core set as otherwise whole subsystems would be included in the output model due to one gene. Nevertheless, the inclusion of barcode-supported genes should be preferred over non-core reaction which are not supported and therefore barcode-supported transporters are not be penalized. For more details on FASTCORE see the original paper [21]. A Matlab implementation of the FASTCORE algorithm can be downloaded from [bio.uni.lu/systems\\_biology/software](http://bio.uni.lu/systems_biology/software). Two optional steps can be included in the workflow. The first one allows to constrain the model with respect to the medium composition if this information is available. The uptake reactions for metabolites that are not present in the medium are shut down and FASTCC [21] is run to remove reactions that cannot carry a flux due to these medium constraints. The second optional step allows adding a biomass function to the model. FASTCORMICS forces the biomass function to carry a flux while penalizing the inclusion of non-core reactions (Figure 1). Core reactions, including core transporters are not penalized in order to find, within the different alternatives sets of reactions that allows the production of biomass, the one that contain the highest number of core reactions. The output reactions of the modified FASTCORE are then added to the core set and the modified FASTCORE is run a second time to now force all the core reactions to carry a flux while penalizing the non-core reactions. Transporter reactions are removed from the core set but are not penalized during the reconstruction to favour barcode-supported transporters over non-core reactions that are not supported. If no biomass function was added, Fastcormics is only run ones.

## 2.2 Microarray preprocessing with Barcode of cancer cell line data

The NCI dataset composed of 174 Hgu133plus2 arrays corresponding to 59 cancer cell lines was downloaded from the Cell miner web page [19] and read in version 2.15.1 of R with the affy package (1.36.1). The arrays were normalized with the frozen Robust Multi-array Average package (fRMA version 1.14.0) [13] and then processed with the Barcode [23] function using the hgu133plus2frmavrecs vector (version 1.1.12) into a list of expressed genes and another list of genes with intensity values not significantly different from the intensities obtained for same probesets in an unexpressed state (Figure 1). The ubiquity of expression (number of arrays for which a gene is expressed over the total number of arrays) was computed for each gene and a list of genes Entrez IDs with their respective score was then loaded in Matlab (version 2013a) and mapped via the Gene Protein Reactions Rules (GPR) to the consistent version of Recon1 (consistRecon1, 2469) and Recon2 (consistRecon2, 5317) obtained with FASTCC. Reactions tagged as expressed in 90% of the 174 arrays were included in the core set with the exception of Barcode-supported transport reactions (core transporters), as the genes controlling the latter are in general promiscuous. The core transporters were excluded from the core set, but were not penalized later during the building process.

## 2.3 Building of cancer models constraint to growth on RPMI medium

To simulate the growth of the cancer cells on RPMI medium, the uptake reactions of the consistent versions of Recon 1 and Recon 2 were first constrained with respect to the medium composition and a biomass function taken from [22] was added to the GEMs. FASTCC [21]

was run to remove reactions that are not able to carry a flux due to these additional medium constraints (Figure 1).

The modified FASTCORE was then run on the medium-constrained models forcing the biomass function to carry a flux while penalizing the inclusion of non-core reactions. This step allows selecting preferentially among all the alternative pathways for the production of biomass the one with the lowest number of non-core reactions, not penalizing the inclusion of core reactions. The reactions required to allow a biomass production were then added to the core set and the modified FASTCORE was run again now forcing the inclusion of all core reaction while penalizing the non-core reactions with the exception of core transporters (Figure 1).

## 2.4 Model validation based on a knock-out experiment to identify essential genes

A knock-out experiment was performed on the obtained cancer models as described by [7] applying Flux Balance Analysis (FBA) [15]. In [7], a gene is considered essential if its knock-downs results in a decrease of the growth rate of more than 1%. To allow, a comparison with [7], the 1% criteria was taken over, but the experiments were also repeated with a growth rate decrease criteria of 50% (growth -50%). The lists of essential genes were compared to the ranked list of 8000 genes established by [10] based on a shRNA knockdown screen on cancer cell lines. The rank of essential genes were compared to the rank remaining metabolic genes (set of genes associated to Recon2 minus the essential genes) with a Kolmogorov-Smirnov test (KS-test). In addition 1000000 random sets of genes of the same size were created and the respective KS-test was computed for evaluating the likelihood to obtain the same or better KS-score by chance. To further validate the predicted essential genes, a list of neoplasia-related genes was retrieved from DisGeNET [16], a database for gene-disease associations. A hypergeometric test was performed to evaluate the enrichment of neoplasia-related genes in the predicted essential genes.

## Results

Two generic cancer models were obtained via the integration of microarray data from the NCI dataset [19] with the FASTCORMICS workflow. The first model (cancer1), derived from Recon 1, is composed of 816 reactions (Table 1) and is therefore bigger than the cancer model derived in [7] (772 reactions). The second model (cancer2) was extracted from Recon 2 and is composed of 1332 reactions. Essentiality assays performed on cancer1 and cancer2 predict 188 and 106 essential genes, respectively. The lists of essential genes were compared to a ranked list of 8000 genes established by [10] via a shRNA knock-outs assay. Metabolic genes are slightly overrepresented in the top of the list (data not shown), suggesting that metabolic genes of Recon 1 and Recon 2 are more essential than the remaining genes on the list. Furthermore, the distribution of essential genes of the cancer models is different from the remaining metabolic genes and shifted towards the top of the ranked list as showed by a one-side KS-test giving p-values of 7.7232e-04 and 0.0095 for cancer1 and cancer2, respectively, compared to a p-value of 0.0284 for the cancer model published in [7] that was built by the MBA algorithm [9].

A permutation test (Table 1) showed that the likelihood of finding a gene set of the same size with a better KS-score by chance is low with a p-value of 1e-6 and 0.0022 for cancer1 and cancer2, respectively, against a p-value of 0.0044 for the cancer model built by the MBA algorithm [7].

The Recon1 and Recon 2 models, further constrained by the medium composition, allowed to only identify a smaller set of essential genes (Table 1) and their distribution were not significantly

**Table 1.** Comparison of the essential genes found by the in silico essentiality assay to a ranked gene list established by [10] defined as the effect of shRNA knock-downs on the proliferation of cancer cells. In [7], a gene is considered essential if its knock-downs resulted in a decrease of the growth rate of more than 1%. To allow, a comparison with [7], the 1% criteria was taken over, but the experiments were also repeated with a growth rate decrease criteria of 50% (growth -50%). \*The number of essential genes was taken from supplementary file msb201135-sup-0002.xls of [7]

output model	generic model	size	essential genes	p-value	permutation p-value
cancer folger	Recon1	772	178*	0.0284	0.0059
medium constrained Recon 1 + biomass	Recon 1	1922	78	0.1908	0.1444
medium constrained Recon 2 + biomass	Recon 2	4246	36	0.7777	0.7398
cancer1	Recon1	816	188	7.7232e-04	1e-06
cancer1 (growth -50 %)	Recon1	816	92	0.0489	0.0324
cancer2	Recon2	1332	106	0.0095	0.0022

**Table 2.** Hypergeometric test quantifying the enrichment of neoplasia-related genes retrieved from DisGeNet [16] in the set of essential genes, a database of disease-gene associations. In [7], a gene is considered essential if its knock-downs resulted in a decrease of the growth rate of more than 1%. To allow, a comparison with [7], the 1% criteria was taken over, but the experiments were also repeated with a growth rate decrease criteria of 50% (growth -50%).

output model	essential genes (EG)	EG in Dis- GeNet	genes in the generic models (GG)	GG in Dis- GeNet	p-value
cancer folger	178	84	1168	449	4.6 e-06
medium constrained Re- con1 + biomass	78	33	1168	377	0.0350
medium constrained Recon2 + biomass	36	14	1599	433	0.0806
cancer1	188	90	1168	377	8.1e-07
cancer1 (growth -50%)	92	39	1168	377	0.0063
cancer2	106	45	1599	433	2.9e-04

different from the distribution of the metabolic genes for Recon1 and Recon 2 , confirming that context-specific models perform better than the generic genome-wide reconstructions from which they are extracted.

The hypergeometric test (Table 2) showed that the neoplasia-associated genes retrieved from the DisGeNet database [16] are over-represented in the essential genes for all models. This confirms that the essential genes predicted by the cancer models are not false positives due to model-specific bias like the lack of alternative pathways in the generic model or the removal of the latter due to a high threshold.

## Discussion

We extended the FASTCORE algorithm [21], that allows reconstructing compact context-specific metabolic models in the time order of seconds, towards the integration of microarray data. The resulting FASTCORMICS workflow, that performs a reconstruction in the order of a few minutes, was validated by an essentiality test performed on generic cancer models, built by FASTCORMICS following the integration of an NCI microarray dataset. The essential genes predicted by FASTCORMICS ranked highly in the list of essential genes established by [10] based on the shRNA knockdown effects on cancer cells proliferation and also were enriched for neoplasia-related genes.

Unlike most competing algorithms, FASTCORMICS does not depend on the introduction of heuristic thresholds for the segregation of expressed and non-expressed genes, which turns the models built by FASTCORMICS more robust and less prompted to over-fitting of the data. For the cancer model built by the competing MBA algorithm, considered as state of the art for context-specific metabolic reconstructions, the threshold was set at an intensity value of 200 [7]. The size of the output model and therefore the number of essential genes is very sensitive to the choice of this threshold. A higher threshold would have led to overestimation of essential genes, due to a reduced number of alternative pathways included in the model, whereas a lower threshold would have led to an underestimation of the number of essential genes. Moreover, as the intensity distribution varies between experiments and platforms, the value of 200 that seems adequate for this dataset, might be inadequate for another one. Further, when comparing cancer1 with the cancer model built by MBA algorithm [9], cancer1 performed slightly better on the essentiality test and the enrichment test. Finally, like for FASTCORE, core reactions fed to MBA must have a high confidence level as all core reactions will be included in the output model. Core reactions with low confidence level, due to a too low threshold or lab effects, can cause the inclusion of a great number of non-core reactions into the output model in order to guarantee a flux through all core reactions. Barcode [23] used with the default setting (as performed here) requires five standard deviation above its null mean to be regarded as expressed which drastically limits the source of error in establishing the core reactions set. Other competing algorithms using omics data for building of context-specific model like GIMME [3], IMAT [24] or, mCADRE [22], INIT [1] or the MBA algorithm [9] have higher computational demands due to the used mixed integer linear programming, and/or require setting of one expression, respectively two expression thresholds for IMAT [24]. FASTCORMICS outperforms its competitors due to its robustness and the low computational demand due to the efficient use of linear programming.

## Acknowledgements

We would like to thank Thomas Pfau and Lasse Sinkkonen for their feedback and for the interesting discussions.

The research was funded by the Life Sciences Research Unit, University of Luxembourg. MPP was supported by a fellowship from the National Research Foundation of Luxembourg (FNR; <http://www.fnr.lu>) (AFR 6041230). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

- [1] R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Computational Biology*, 8(5):e1002518, 2012.
- [2] Rasmus Agren, Adil Mardinoglu, Anna Asplund, Caroline Kampf, Mathias Uhlen, and Jens Nielsen. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular systems biology*, 10(3), 2014.
- [3] S. A. Becker and B. Ø. Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, 2008.
- [4] A. Bordbar, N. E. Lewis, J. Schellenberger, B. Ø. Palsson, and N. Jamshidi. Insight into human alveolar macrophage and m. tuberculosis interactions via metabolic reconstructions. *Molecular systems biology*, 6(1), 2010.
- [5] Ron Caspi, Tomer Altman, Joseph M Dale, Kate Dreher, Carol A Fulcher, Fred Gilham, Pallavi Kaipa, Athikkattuvalasu S Karthikeyan, Anamika Kothari, Markus Krummenacker, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 38(suppl 1):D473–D479, 2010.
- [6] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007.
- [7] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppín, and T. Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology*, 7(501), 2011.
- [8] Daniel R Hyduke, Nathan E Lewis, and Bernhard Ø Palsson. Analysis of omics data with genome-scale models of metabolism. *Mol. BioSyst.*, 9(2):167–174, 2013.
- [9] L. Jerby, T. Shlomi, and E. Ruppín. Computational reconstruction of tissue-specific metabolic models: Application to human liver metabolism. *Molecular Systems Biology*, 6(401), 2010.
- [10] Biao Luo, Hiu Wing Cheung, Aravind Subramanian, Tanaz Sharifnia, Michael Okamoto, Xiaoping Yang, Greg Hinkle, Jesse S Boehm, Rameen Beroukhim, Barbara A Weir, et al. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences*, 105(51):20380–20385, 2008.
- [11] Hongwu Ma, Anatoly Sorokin, Alexander Mazein, Alex Selkov, Evgeni Selkov, Oleg Demin, and Igor Goryanin. The edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135, 2007.



- [12] Adil Mardinoglu, Rasmus Agren, Caroline Kampf, Anna Asplund, Mathias Uhlen, and Jens Nielsen. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications*, 5, 2014.
- [13] Matthew N McCall, Benjamin M Bolstad, and Rafael A Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010.
- [14] Matthew N McCall, Karan Uppal, Harris A Jaffee, Michael J Zilliox, and Rafael A Irizarry. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research*, 39(suppl 1):D1011–D1015, 2011.
- [15] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism. *Molecular systems biology*, 7(1), 2011.
- [16] Núria Queralt-Rosinach and Laura Inés Furlong. Disgenet rdf: A gene-disease association linked open data resource. In *SWAT4LS*, 2013.
- [17] Pedro Romero, Jonathan Wagg, Michelle L Green, Dale Kaiser, Markus Krummenacker, and Peter D Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, 6(1):R2, 2004.
- [18] Daniel Segre, Dennis Vitkup, and George M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, 2002.
- [19] Uma T Shankavaram, Sudhir Varma, David Kane, Margot Sunshine, Krishna K Chary, William C Reinhold, Yves Pommier, and John N Weinstein. Cellminer: a relational database and query tool for the nci-60 cancer cell lines. *BMC genomics*, 10(1):277, 2009.
- [20] I. Thiele, N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, et al. A community-driven global reconstruction of human metabolism. *Nature biotechnology* (doi:10.1038/nbt.2488), 2013. doi:10.1038/nbt.2488.
- [21] Nikos Vlassis, Maria Pires Pacheco, and Thomas Sauter. Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput Biol*, 10(1):e1003424, 01 2014.
- [22] Y. Wang, J. A. Eddy, and N. D. Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Systems Biology*, 6(1):153, 2012.
- [23] Michael J Zilliox and Rafael A Irizarry. A gene expression bar code for microarray data. *Nat Meth*, 4(11):911–913, November 2007.
- [24] Hadas Zur, Eytan Ruppin, and Tomer Shlomi. imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24):3140–3142, 2010.